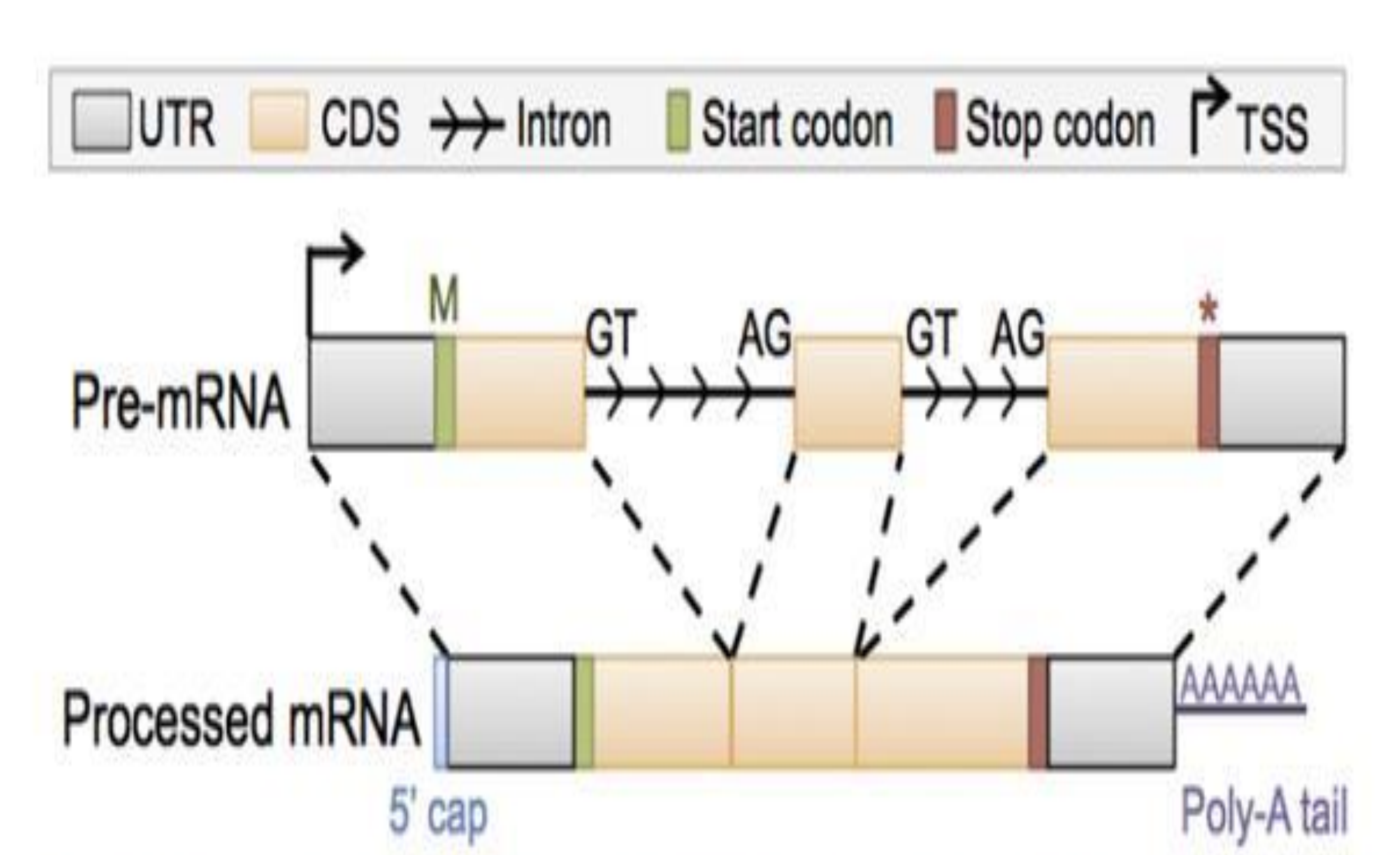
Annotation of the Genome of *Drosophila ananassae*

Landon Kehr and Dr. Sarah Justice

BACKGROUND

- The goal of this project is to annotate a sequence of the genome of the *Drosophila ananassae* fruit fly species
- The genes in this sequence are located on a chromosomal region known as the Muller D-element.
- This region is composed mainly of a type of genetic material called euchromatin which is loosely packed and more heavily expressed as genes than its counterpart heterochromatin.
- Annotating these genes will hopefully help us better understand how variations between species in heavily expressed areas of DNA can still produce viable individuals as these mutations are often fatal.
- Comparing the genomes of *D. ananassae* and *D. melanogaster* will hopefully give us more insight into how mutations in euchromatic DNA could affect humans.



Typical Gene Structure with characteristics of the genes we are trying to identify shown.

METHODS

1. Using the UCSC Genome Browser and Genscan gene predictions of the targeted region of the *D. ananassae* genome, I obtained the predicted protein sequence based on the sequence of nucleotides.

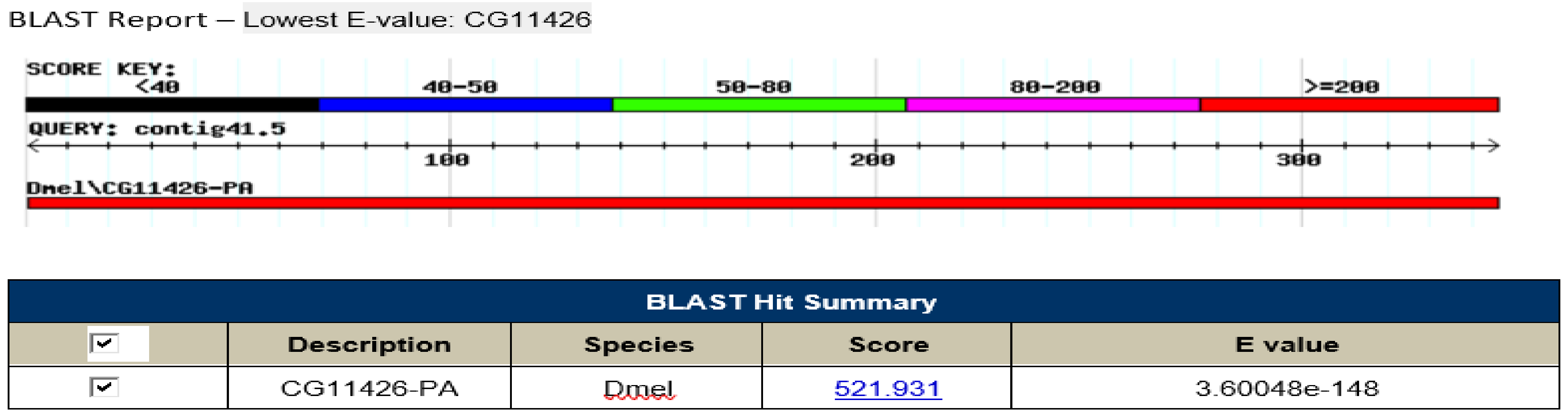
2. I then used the Basic Local Alignment Search Tool (BLAST) to find the analogous gene in the *D. melanogaster* species.

3. Next, I used the FlyBase Gene Record Finder to find the predicted start and end sites as well as any potential splice sites and the strand of DNA on which the gene is encoded.

4. Finally, I returned to the UCSC browser and examined the DNA sequence to determine if the start, end, and splice sites that were predicted are feasible.

RESULTS

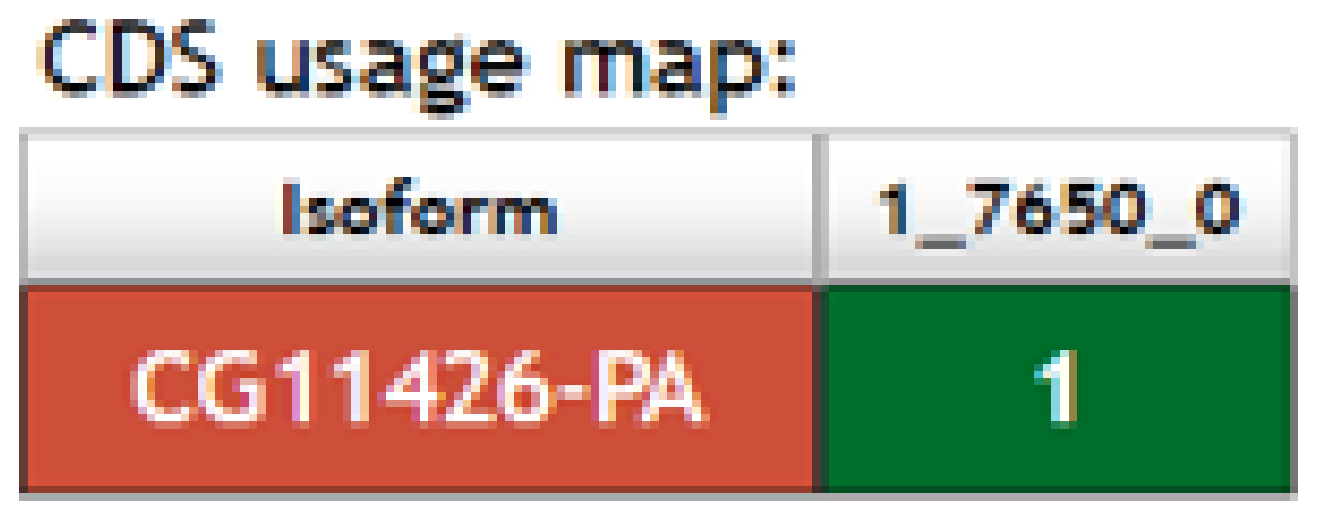
Using BLAST, I found that the Genscan predicted protein sequence is most closely aligned to the CG11426 gene in *Drosophila melanogaster*. The E Value shown below is close to 0, indicating a strong alignment.



BLAST Report – Lowest E-value: CG11426

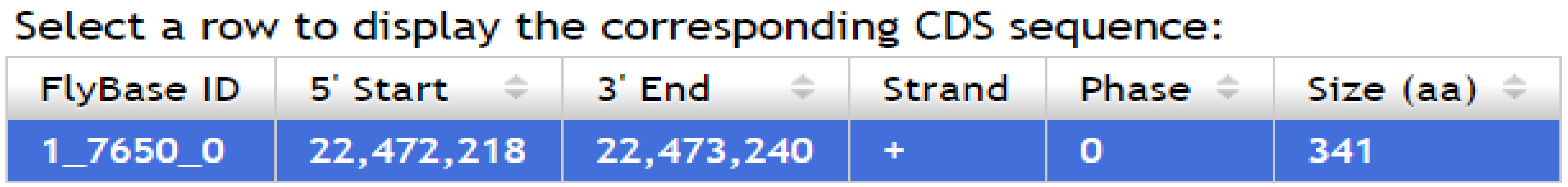
BLAST Hit Summary				
☑	Description	Species	Score	E value
☑	CG11426-PA	Dmel	521.931	3.60048e-148

The FlyBase Gene Record Finder revealed that the CG11426 gene has only 1 isoform with only 1 exon (coding sequence). It also predicts that the gene is on the positive DNA strand meaning it will be translated from left to right across the genome browser.



CDS usage map:

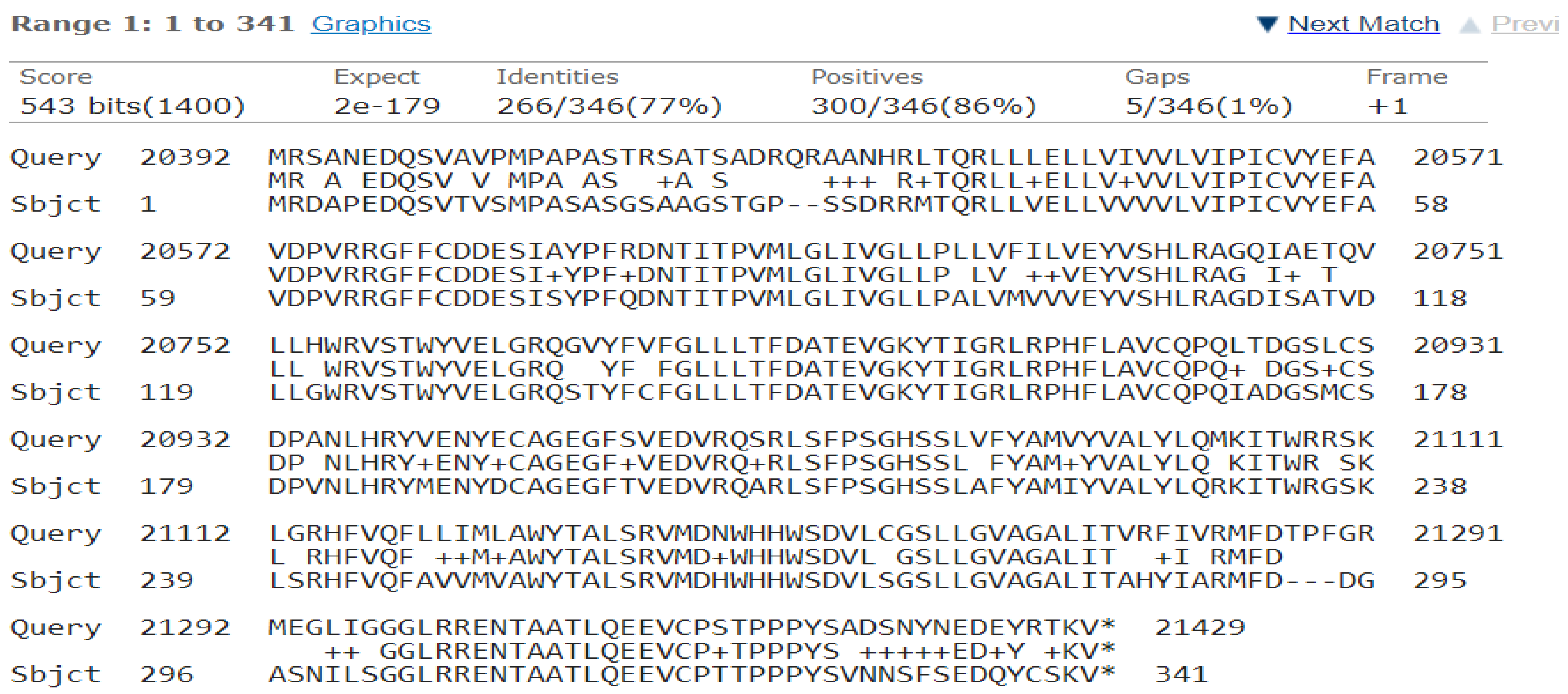
Isoform	1_7650_0
CG11426-PA	1



Select a row to display the corresponding CDS sequence:

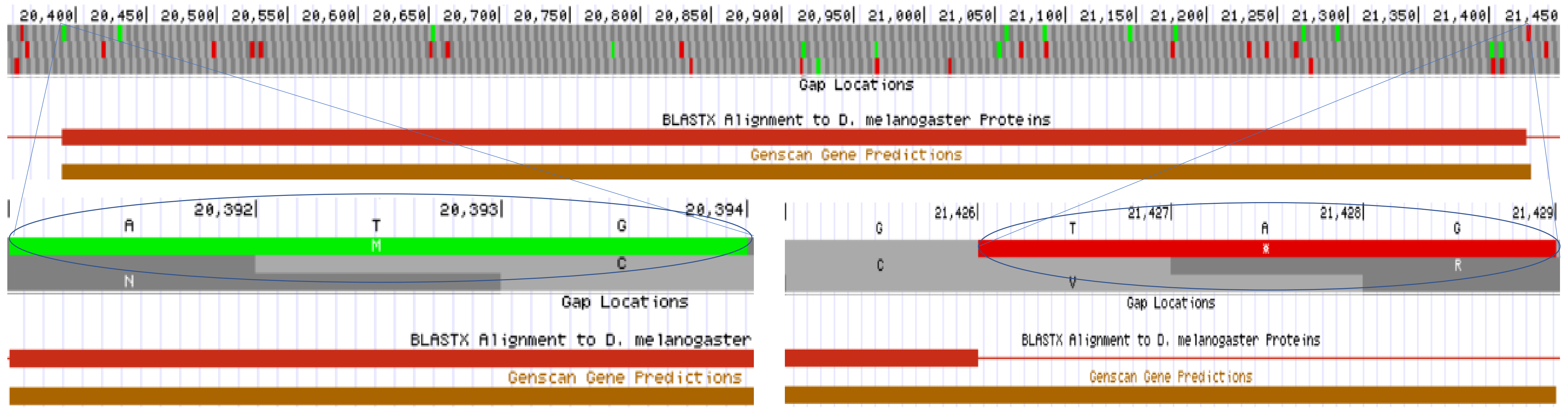
FlyBase ID	5' Start	3' End	Strand	Phase	Size (aa)
1_7650_0	22,472,218	22,473,240	+	0	341

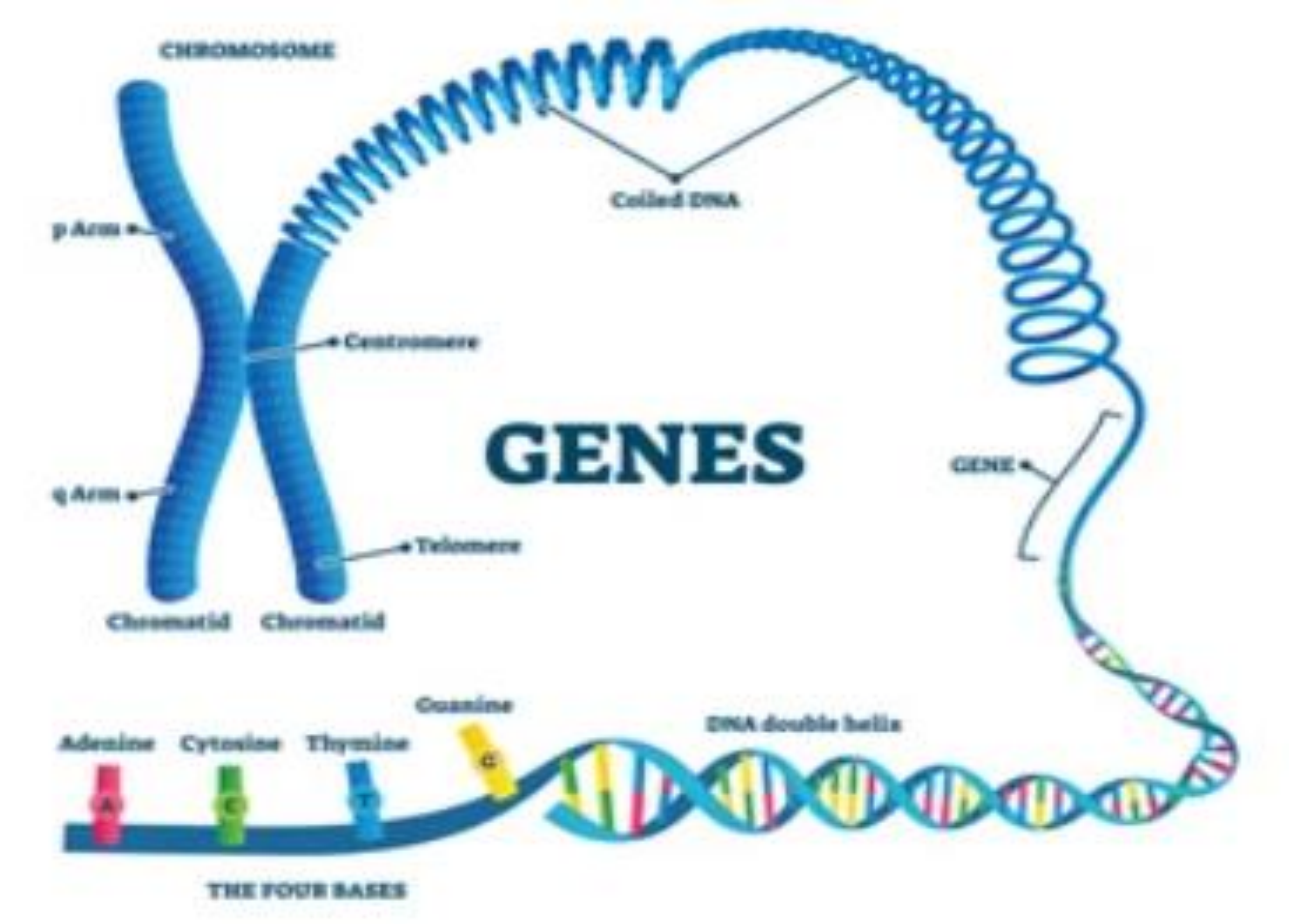
The BLAST alignment shown below predicted that the analogue in *D. ananassae* starts at nucleotide 20,392 and ends at nucleotide 21,429 and is located in frame +1.

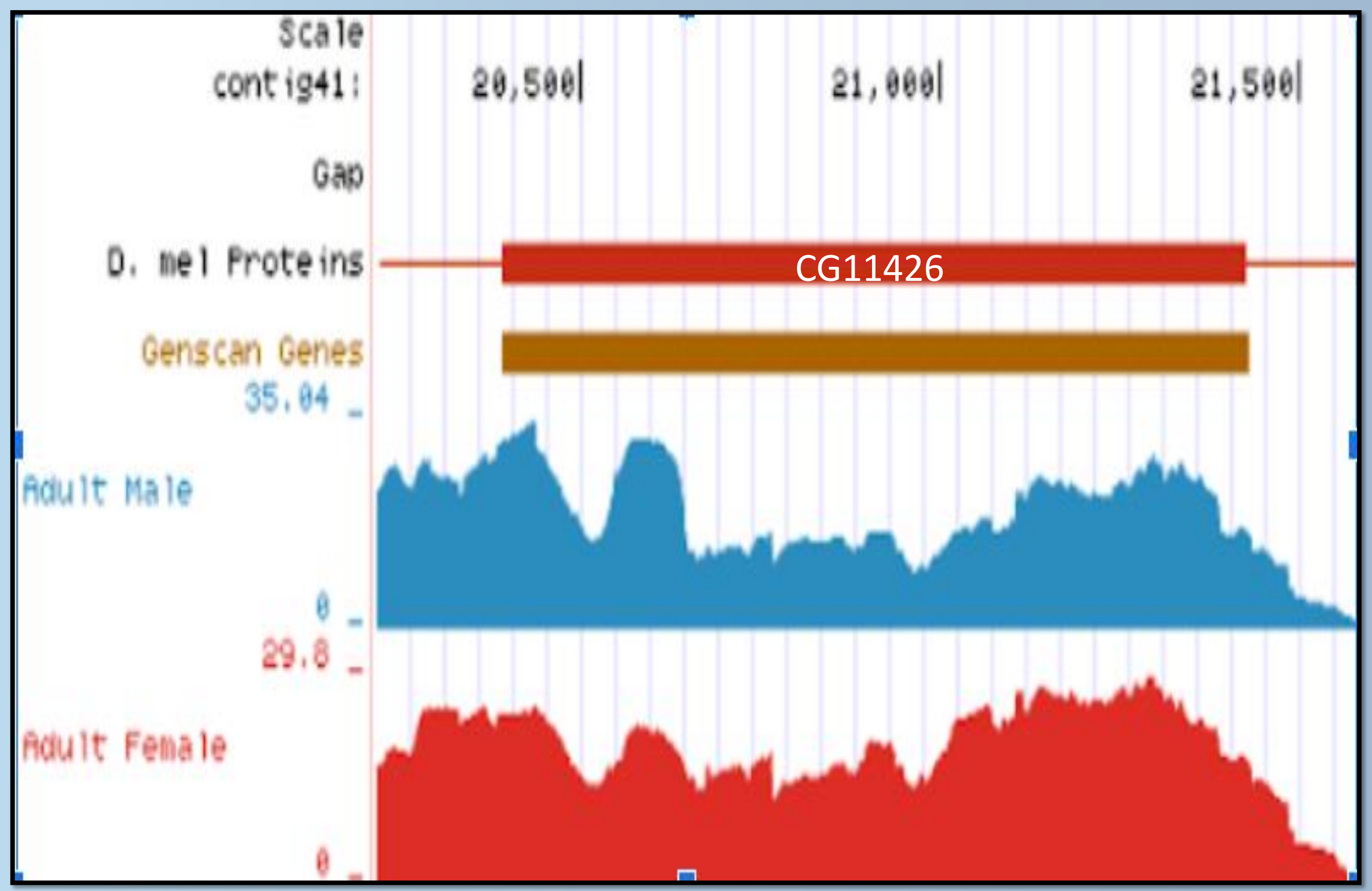


```
Range 1: 1 to 341  Graphics                                    ▼ Next Match  ▲ Previ

Score           Expect    Identities      Positives       Gaps        Frame
543 bits(1400)  2e-179    266/346(77%)    300/346(86%)    5/346(1%)   +1

Query  20392  MRSANEDQSVAVPMPAPASTRSATSADRQRAANHRLTQRLLLELLVIVVLVIPICVYEFA  20571
              MR A EDQSV V MPA AS  +A S     +++ R+TQRLL+ELLV+VVLVIPICVYEFA
Sbjct  1      MRDAPEDQSVTVSMPASASGSAAGSTGP--SSDRRMTQRLLVELLVVVVLVIPICVYEFA  58

Query  20572  VDPVRRGFFCDDESIAYPFRDNTITPVMLGLIVGLLPLLVFILVEYVSHLRAGQIAETQV  20751
              VDPVRRGFFCDDESI+YPF+DNTITPVMLGLIVGLLP LV ++VEYVSHLRAG I+ T
Sbjct  59     VDPVRRGFFCDDESISYPFQDNTITPVMLGLIVGLLPALVMVVVEYVSHLRAGDISATVD  118

Query  20752  LLHWRVSTWYVELGRQGVYFVFGLLLTFDATEVGKYTIGRLRPHFLAVCQPQLTDGSLCS  20931
              LL WRVSTWYVELGRQ  YF FGLLLTFDATEVGKYTIGRLRPHFLAVCQPQ+ DGS+CS
Sbjct  119    LLGWRVSTWYVELGRQSTYFCFGLLLTFDATEVGKYTIGRLRPHFLAVCQPQIADGSMCS  178

Query  20932  DPANLHRYVENYECAGEGFSVEDVRQSRLSFPSGHSSLVFYAMVYVALYLQMKITWRRSK  21111
              DP NLHRY+ENY+CAGEGF+VEDVRQ+RLSFPSGHSSL FYAM+YVALYLQ KITWR SK
Sbjct  179    DPVNLHRYMENYDCAGEGFTVEDVRQARLSFPSGHSSLAFYAMIYVALYLQRKITWRGSK  238

Query  21112  LGRHFVQFLLIMLAWYTALSRVMDNWHHWSDVLCGSLLGVAGALITVRFIVRMFDTPFGR  21291
              L RHFVQF ++M+AWYTALSRVMD+WHHWSDVL GSLLGVAGALIT  +I RMFD
Sbjct  239    LSRHFVQFAVVMVAWYTALSRVMDHWHHWSDVLSGSLLGVAGALITAHYIARMFD---DG  295

Query  21292  MEGLIGGGLRRENTAATLQEEVCPSTPPPYSADSNYNEDEYRTKV*  21429
                 ++ GGLRRENTAATLQEEVCP+TPPPYS +++++ED+Y +KV*
Sbjct  296    ASNILSGGLRRENTAATLQEEVCPTTPPPYSVNNSFSEDQYCSKV*  341
```

This was confirmed using the genome browser (shown below). In frame +1 on the positive DNA strand (the top row) there is a start codon (green) at nucleotides 20,392-20,394 and a stop codon (red) at nucleotides 21,427-21,429. There are also no stop codons in this frame that would interrupt the translation.







A section of the UCSC Genome Browser used, showing the location in the genome of *D. ananassae*, known *D. melanogaster* genes (red), predicted Genscan Genes (brown), and RNA Sequencing data (blue and red graphs).

CONCLUSIONS

In conclusion our data suggests that the region of the genome of *Drosophila ananassae* corresponds to the CG11426 gene in *Drosophila melanogaster.* Moreover, the abundant RNA sequencing data indicates that this region of the genome is often transcribed in both male and female adults. The next stage of the project will include locating a transcription start site and continuing to identify analogous genes in the area.

ACKNOWLEDGEMENTS

- This project was done in collaboration with the Genomics Education Partnership.
- Special thanks to Dr. Sarah Justice and the members of my research group Kayla Britt, Hailey Chang, Rebecca Hatcher, Sam Nolan, Grace Reynolds, and Tia Watkins

REFERENCES

Picture Credits:
https://www.shutterstock.com/search/chromosome
https://www.insidehighered.com/sites/default/server_files/media/24129965_10156071642798648_7524652002160677436_n.png
Background Information:
https://flybase.org/maps/synteny
https://thegep.org/felement/